Small Pixels

ビデオ符号化高度化

ソリューション概要

ホワイトペーパー（和訳）

Document ID

WHITEPAPER - VIDEO CODING

Date

19/02/2026

Version

1.0_JP

和訳

株式会社ヴィレッジアイランド

目次

Small Pixels

ビデオ符号化高度化

ソリューション概要

● 概要（ABSTRACT）

本ホワイトペーパーは、Small Pixels 社が提供する映像符号化高度化（Video Coding Enhancement）ソリューションについて解説するものである。新規導入を検討する技術者に対して包括的な理解を提供するとともに、ニューラルネットワークを活用した映像符号化最適化技術の技術的背景を詳述する。

想定読者は、CTO、アーキテクト、開発者、運用担当者など、技術的意思決定に関与する専門家である。本書を通じて、Deep Pre-Editing（ディープ事前編集）という概念に基づく Small Pixels 社のアプローチの中核技術を理解できる。

1. はじめに

本書では、ニューラルネットワークを活用した映像符号化高度化技術の概要、および Small Pixels 社の提供するソリューションの基本思想を紹介する。

第 1 章では NN ベース符号化高度化の概念を概観し、第 2 章で技術的実装および最適化理論を解説し、第 3 章で結論と将来展望を示す。

1.1 Deep Pre-Editing による高度映像符号化

帯域が制約されたネットワーク上で高忠実度の映像データを伝送することは、現代のデジタル通信における根本的な課題である。歴史的に、通信業界およびストリーミング業界は、標準的な映像コーデックの継続的な進化に依存してきた。代表的なものとして、広く普及している Advanced Video Coding（AVC/H.264）、High Efficiency Video Coding（HEVC/H.265）、Versatile Video Coding（VVC/H.266）、そしてオープンソースの AV1 が挙げられる。これらのアルゴリズムは、高度に最適化されたレート-歪み（Rate-Distortion）最適化機構として機能する。

これらのコーデックは、映像ストリームを Coding Tree Unit（CTU）やマクロブロックへと構造的に分解し、フレーム内空間予測およびフレーム間動き補償を用いて残差信号を抽出する。抽出された残差は通常、離散コサイン変換（DCT）によって周波数領域へ変換され、厳密に定義されたビット予算内に収めるために量子化される［DAS-2023］。

これらの標準規格は高度に洗練された設計を有しているものの、シーンの複雑さに対して割り当てビットレートが不足した場合、量子化処理によって変換係数が不可避的に切り捨てられるという本質的制約に依存している。この積極的な切り捨ては映像信号から高周波成分のエネルギーを除去する。人間の視覚系にとって、この高周波情報の欠落は重大な視覚アーティファクトとして知覚される。具体的には、マクロブロック境界におけるブロッキング、コントラストの高いエッジ周辺のモスキートノイズ、色再現性の劣化、全体的なぼやけなどが挙げられる［BERTINI-2022, TALEBI-2021］。

これまで、これらのアーティファクトを軽減する標準的アプローチは、クライアント側でポストプロセッシングフィルタを適用することであった。デコード後段で動作するアルゴリズムは、ブロック境界を平滑化し、失われたテクスチャを推定的に補完（ハルシネーション）する。しかしポストプロセッシングは本質的に情報ボトルネックの制約を受ける。すなわち、数学的に劣化した信号から元の現実を再構築しなければならないからである。

Small Pixels のアーキテクチャは、この問題に対して全く異なるアプローチを採用している。それが「Deep Pre-Editing（ディープ事前編集）」と呼ばれるパラダイムである［TALEBI-2021］。この中核概念は、もともと Hossein Talebi らによって JPEG 静止画向けに提案されたものであり、エンコード前にソース信号を知的に改変することで、低ビットレート環境において下流のコーデックがより有利な動作領域で機能できるようにするというものである。その結果、同等の知覚品質において可視アーティファクトを削減する、あるいはより低いビットレートを実現することが可能となる。

Small Pixels エンジンでは、学習済みニューラルネットワーク（NN）をエンコード前段に配置し、標準コーデックに入力される前の映像フレームを知的に変換する。この

セマンティックおよび空間的適応処理により、映像の固有エントロピーが能動的に低減され、コンテンツは本質的に圧縮しやすい構造へと変換される。

次章では、Deep Pre-Editing パラダイムの技術的解析を行い、エンドツーエンド学習を可能にする数理的定式化、関与するニューラルネットワークのアーキテクチャ構造、微分可能コーデックプロキシの実装、そして全体プロセスの直観的理解を提示する。

Small Pixels の手法は、映像コンテンツを知的に改変することにより、従来より優れた映像符号化を実現する。この戦略的編集は、以下の三つの主要目的を同時に達成するよう設計されている。

1. 視覚情報の大部分を保持すること
2. 後段の圧縮プロセスを最適化し、アーティファクトを大幅に低減すること
3. 圧縮後も視覚的に自然で魅力的な映像を維持すること

これらの要素を統合することで、指定されたビット予算内における圧縮-復号サイクルの性能を最大化する革新的ソリューションが実現されている。

1.2 映像圧縮における Rate-Distortion-Perception (RDP) ボトルネック

Deep Pre-Editing [TALEBI-2021] の必要性および有効性を理解するためには、従来のレート-歪み (Rate-Distortion, R-D) 理論の限界を評価する必要がある。標準的な映像エンコーダの設計は、厳格なビットレート制約のもとで、客観的歪み指標の最小化に基づいている。歪みは通常、Mean Squared Error (MSE) や Peak Signal-to-Noise Ratio (PSNR) といったピクセル単位の指標によって評価される。エンコーダは、与えられたビット数の範囲内で最小の MSE を実現するため、ブロック分割方式、動きベクトル、量子化パラメータの最適な組み合わせを継続的に探索する。

しかし近年、情報理論および視覚データ圧縮におけるディープラーニングの進展により、Rate-Distortion-Perception (RDP) トレードオフが理論的に定式化された。この原理は、歪みを厳密に最小化すること (PSNR を最大化すること) が、知覚品質の最適化と本質的に対立する場合が多いことを数学的に示している。MSE はピクセル単位の誤差のみを評価するため、人間の視覚系が意味構造、テクスチャ、時間的連続性を総合的に解釈する複雑なプロセスを完全に無視している [KHAN-2025, DING-2021]。ディープラーニングに基づく学習目的関数は、歪み指標 (PSNR/SSIM) の最適化と、MOS・VMAF・LPIPS などの知覚指標による「知覚的妥当性」との間に存在する緊張関係を適切に扱わなければならない。この知覚-歪みトレードオフは [BLAU-2018] により理論化されている。

HEVC や VVC のような標準コーデックが極端にビット不足の状態に置かれた場合、その設計原理に従い、グローバル MSE を最小化しようとする。水面の不規則な揺らぎ、フィルムグレイン、森林の複雑な葉の構造といった確率的かつ高周波なテクスチャは、

正確に符号化するために膨大なビットを必要とする。そのため、コーデックは量子化行列によりこれらの高周波係数をゼロ化する。結果として、限られたビット予算はピクセルブロックの数学的平均値を維持することに費やされる。こうして得られる復号フレームは PSNR 上は最適であっても、視覚的には平板で人工的（プラスチック的）に見え、硬直したブロックアーティファクトが顕著に現れるという、知覚的には破綻した映像となる。

Deep Pre-Editing は、この RDP トレードオフを意図的に活用することで機能する。Small Pixels の事前編集エンジンは、非圧縮の高品質な原映像に対して、アルゴリズム的に制御された知覚的に不可視な歪みをあえて導入する。この事前歪みは、カメラセンサーデータとの厳密なピクセル一致という数学的忠実性を犠牲にする代わりに、後段の標準コーデックが最大効率で処理可能な構造的特性を付与する。その結果、コーデック自身の仕様上不可避である詳細除去に伴う実際の歪みを低減することができる。

知覚的に重要でない領域に対して知的に作用することで、事前編集器はフレームの空間分散（spatial variance）を能動的に低減する。その結果、標準エンコーダはより緩和された数理的動作領域で処理を行うことが可能となる。これにより、意味的に重要な構造部分に対してより細かな量子化ステップを適用できるようになる。最終的に得られる圧縮-復号後の映像は、視覚品質および人間の知覚において大幅に優れたものとなる。

Small Pixels のニューラルネットワークは、実質的にエンコーダに対する「プラグイン」のように機能し、従来実装の単純な損失圧縮パラメータを、より高度で知的な NN ベースのソリューションへと置き換えていると言える。

2. ニューラルネットワークによる映像符号化高度化

NN ベースの映像符号化高度化とは、特定用途向けに学習されたニューラルネットワークを用いて、対象コンテンツの映像符号化（すなわち映像圧縮）性能を向上させる技術である。

本方式では、高品質な映像を入力とし、人間の視覚にとって知覚的にほぼ同等でありながら、より圧縮しやすい特性を持つ別の高品質映像を出力する。その結果、最終的な符号化映像は、ユーザーにとってより高品質に知覚されると同時に、より高い圧縮効率を実現する。

一般的なユースケースにおいては、出力映像の解像度およびフレームレートは入力と完全に同一に維持される。ニューラルネットワークの目的は、後段に配置される標準ビデオエンコーダの性能を最適化することである。

入力映像はフレーム単位で処理される。この構成により、システム全体の遅延は実質的に 1 フレームに抑えられるため、ライブ配信および VOD（Video on Demand）エンコードの双方に適用可能である [GALTERI-2019, VACCARO-2021]。

各フレームはまずニューラルネットワークによって高度化処理が施され、その後ビデオエンコーダに渡される。エンコーダは最終的な高度化済み出力映像を生成する。この一連の処理フローは図 1 に要約されている。

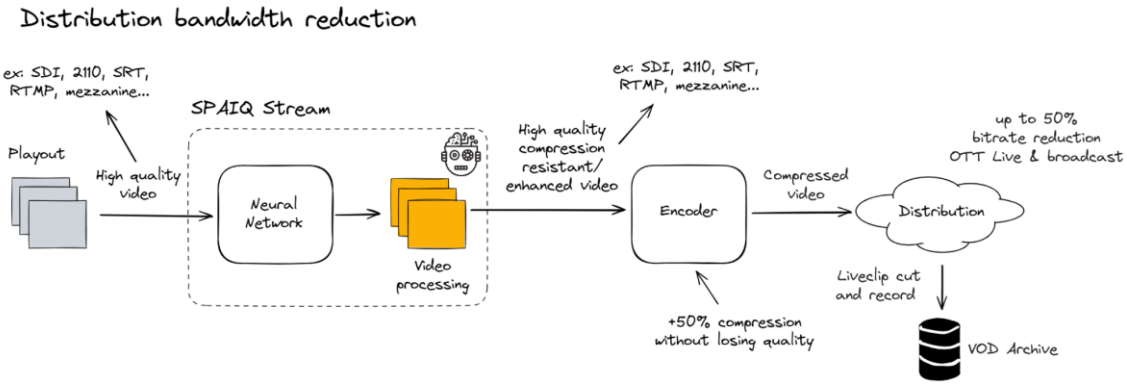

図 1：NN ベース映像符号化高度化プロセス

本映像符号化高度化システムは、標準エンコーダの前段で動作する。この構成により、標準ビットストリームはそのまま維持され、既存のすべてのクライアント環境に変更を加える必要がない。

また、本方式は既存のトランスコード／パッケージングパイプラインへ容易に組み込むことが可能である。Small Pixels のソリューションは、「クライアント側の変更を一切必要としない」ことを中核的価値として設計されている。

さらに、強化ネットワークは一度学習・本番環境へ展開された後は、追加の再学習や設定変更を必要としない。

2.1 最適化問題の定式化

Small Pixels の Deep Pre-Editing パイプラインにおける工学的目的は、任意に入力される映像ストリームに対して自動的にコンテンツ適応処理を実行できる、汎用的なフィードフォワード型ディープニューラルネットワークを構築することである。この処理は、標準エンコードが適用される前段において、サーバ側でリアルタイムに実行されなければならない。

この動作を形式的に定義する。時刻インデックス t における高品質・非圧縮の映像フレームを z_t ∈ R^(H×W×C) とする。ここで z_t は高さ H、幅 W、色チャネル数 C（通常は RGB または YUV）を持つフレームを表す。

標準映像符号化処理（ビットストリームへの圧縮およびその後のピクセル空間への復号を含む）を C_B(・)
という関数で表す。ここで B は厳格なビット予算制約を意味する。
従来の伝送パイプラインでは、復号フレームは x'_t = C_B(z_t)
として得られ、そこにはコーデック固有の圧縮アーティファクトが含まれる。

Deep Pre-Editing では、パラメトリックなニューラル編集ネットワーク T(θ, z_t)
を導入する。ここで θ はネットワークの学習済み重みおよびバイアスを包含する。
ネットワークは元のフレームを入力し、事前編集済みフレーム x_t = T(θ, z_t)
を出力する。
数学的目的は、最終的な圧縮-復号フレーム C_B(x_t)が以下の三つの制約を同時に厳密に満たすように、パラメータ θ を最適化することである。

1. セマンティック近接性
   事前編集フレームは元フレーム z_t と高い意味的および視覚的類似性を維持しなければならない。ネットワークが全く新しい内容を生成する（ハルシネーションする）ことを防ぐ。

2. ビットレート圧縮性
   事前編集フレーム x_t は、元フレーム z_t よりも少ないビットで符号化できる必要がある。あるいは、同一ビット予算 B においてより高い知覚品質を実現しなければならない。
3. 視覚アーティファクト抑制
   最終出力 C_B(x_t) は、より優れた知覚品質を示さなければならない。特に、ブロッキング、リンギング、色にじみなどの除去を目的とする。

未圧縮映像の包括的データセットを用いてネットワーク θ を学習させるための最適化損失関数の定式化は、これらの要件を満たすよう設計されている。その主な構成要素は以下の通りである。

● C'_q(・)
特定の量子化パラメータまたは品質係数 q で動作する対象ビデオコーデックの微分可能プロキシを表す。標準コーデックはバックプロパゲーションの勾配を伝播できないため、微分可能なプロキシは必須である。

● dist(・, ・)

元フレームと圧縮後フレームとの近接度を測る指標である。単純な L1 や L2 のピクセル距離では不十分であり、これらは画像品質を過度に罰則化し、結果としてぼやけた画像を生じさせる。

その代わりに、Small Pixels の実装では、特徴抽出に基づくディープ・パーセプチュアルロスと、Learned Perceptual Image Patch Similarity（LPIPS）[ZHANG-2018] および Video Multimethod Assessment Fusion（VMAF）[LI-2025] といった高度な微分可能知覚指標を組み合わせて用いている。

このように慎重に設計された損失関数は、出力をブロック状やぼやけた状態から継続的に遠ざけ、美的に優れた高品質な再構成へと導く。

2.2 非微分コーデックの障壁の克服

エンドツーエンド型の Deep Pre-Editing ネットワークを学習させるうえで最も大きな障害となるのは、最適化損失関数の内部に標準ビデオコーデック C(·)C(¥cdot)C(·) を必然的に含まなければならない点である。ディープラーニングは、微積分の連鎖律に基づいて勾配を計算するバックプロパゲーションに全面的に依存している。しかし標準ビデオコーデックは、離散的な量子化およびエントロピー符号化を中核としているため、本質的に非微分系である。

量子化とは、連続値の変換係数を量子化行列で除算し、その結果を最も近い整数へ丸める数学的処理である。丸め演算 [x] は階段関数として振る舞う。階段関数の数学的導関数は、段差の境界点を除き、ほぼすべての領域でゼロとなり、その境界では未定義となる。勾配がゼロであるため、バックプロパゲーションは量子化段階で完全に停止してしまう。もし品質評価指標からの勾配がコーデックを通じて事前編集ネットワークへ逆伝播できなければ、CNN は重みを更新できず、自身の空間変形や平滑化処理が最終的な圧縮出力にどのような影響を与えるかを学習することができない。

この数学的な障壁を解決するには、ビデオコーデックの完全に微分可能なプロキシモデルを構築する必要がある。これらはオフライン学習段階でのみ使用される。この特殊な構成要素は、Small Pixels が開発した学習手法において最も重要な要素の一つであり、ブロックベースの動き補償予測 [CHADHA-2021] など、ビデオコーデック特有のアーティファクトや処理要素を再現する微分可能プロキシとともに実装されている。

2.3 高度な微分可能プロキシコーデック

イントラフレーム符号化アーキテクチャにおいて、微分可能プロキシは、標準的な変換符号化の機能を正確に再現しつつ、ゼロでない勾配の連続的な流れを維持しなければならない。色空間変換（RGB から YUV）および変換処理（順方向 DCT および逆方向

DCT）は、本質的に行列演算による線形処理であり、明確かつ容易に計算可能な導関数を有している。
しかしながら、依然として最も重要なボトルネックは量子化段階である。Small Pixels の学習アーキテクチャでは、最新の微分可能コーデック近似技術を活用し、あらゆる圧縮率の範囲において実際のコーデックとの厳密な整合性を確保している。近年の微分可能 JPEG モデルに関する研究［REICH-2024］で示されている通り、従来の近似手法では重要な離散化処理や境界制限を十分にモデル化できず、極端な圧縮レベルにおいて勾配の伝播が著しく悪化することがある。
堅牢な微分可能パイプラインを実現するために、当該プロキシには以下の高度な機構が組み込まれている。

● Differentiable Clipping（微分可能クリッピング）
ピクセル空間におけるコーデック画像および量子化テーブルの双方に対して微分可能なクリッピング処理を適用し、色域の極端領域における勾配爆発を防止する。
● Differentiable Flooring（微分可能フロアリング）
量子化テーブルのスケールおよび量子化行列値そのものに対して微分可能なフロアリング処理を導入する。

● Straight-Through Estimator（STE）
単純な多項式近似のみに依存するのではなく、より堅牢なモデルでは Straight-Through Estimator を用いる。順伝播（フォワードパス）では、標準コーデックが要求する正確な離散丸め処理を実行し、損失計算の前向き評価を厳密に一致させる。逆伝播（バックワードパス）では、丸め関数のゼロ導関数を回避し、あたかも恒等写像であるかのように勾配をそのまま透過させる。

これらの境界条件および離散化処理を完全に微分可能な枠組みでモデル化することにより、本プロキシは品質全域にわたって標準コーデックを極めて高精度に近似することが可能となる。その結果、事前編集ネットワークは下流コーデックの挙動の微細な特性まで正確に学習することが保証される。

3. 結論および今後の展望

標準映像圧縮技術の発展は、もはや明確な収穫逓減の段階に到達している。標準コーデックに対する反復的な改良によって得られるレート-歪み（Rate-Distortion）性能の漸進的向上は、ネットワーク帯域という物理的制約を根本的に克服するには不十分である。
Deep Pre-Editing は、デジタル映像配信パイプラインにおける基盤的かつ不可欠な進化を示すものである。ディープニューラルネットワークを用いて、ソースコンテンツを能動的かつ知覚的に不可

視な形で知的に改変することにより、すなわちニューラルネットワークのセマンティック理解能力を活用して圧縮困難な確率的ノイズを低減し、複雑な幾何構造をブロックベース変換に最適に整合させることで、本パラダイムはエンコーダに入力される生データのエントロピーを本質的に削減する。

知覚品質指標と高度かつ高精度な微分可能コーデックプロキシを組み合わせたエンドツーエンド学習により、これらの事前編集ネットワークは Rate-Distortion-Perception トレードオフを効果的に活用することを学習する。その結果、低ビットレート量子化によって生じる最も過酷で視覚的に破壊的な影響から映像コンテンツを積極的に保護することが可能となる。

重要なのは、この高度な品質向上処理が標準エンコード工程の前段で完全に実行されるため、既存のデコード用ハードウェアインフラストラクチャに一切の変更を必要としない点である。これにより、世界中に存在する数十億台規模の既存エッジデバイスに対して即時の後方互換性が保証される。

参考文献（REFERENCES）

[BERTINI-2022] BERTINI, Marco, GALTERI, Leonaro, SEIDENARI, Lorenzo, URICCHIO, Tiberio, & DEL BIMBO, Alberto. Fast and effective AI approaches for video quality improvement. In: Proceedings of the 1st Mile-High Video Conference. 2022

[BLAU-2018] BLAU, Yochai; MICHAELI, Tomer. The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018

[CHADHA-2021] CHADHA, Aaron; ANDREOPOULOS, Yiannis. Deep perceptual preprocessing for video coding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. p. 14852-14861, 2021

[DAS-2023] DAS, Tanni; CHOI, Kiho; CHOI, Jaeyoung. High quality video frames from VVC: A deep neural network approach. *IEEE Access*, 2023, 11: 54254-54264.

[DING-2021] DING, D., MA, Z., CHEN, D., CHEN, Q., LIU, Z., & ZHU, F. Advances in video compression system using deep neural network: A review and case studies. Proceedings of the IEEE, 109(9), 1494-1520, 2021.

[GALTERI-2019] GALTERI, L., SEIDENARI, L., BERTINI, M., DEL BIMBO, A. (2019). Towards Real-Time Image Enhancement GANs. In: Proc. of Computer Analysis of Images and Patterns (CAIP), 2019

[KHAN-2025] KHAN,Muhammad Umar Karim, CHADHA,Aaron, ANAM,Mohammad Ashraful, ANDREOPULOS,Yiannis. Perceptual Video Compression with Neural Wrapping Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

[LI-2025] LI,J., ZHOU,C., CHEN,Y. and LU,G., Differentiable VMAF: A trainable metric for optimizing video compression codec, In Proc. of IEEE International Symposium on Circuits and Systems (ISCAS),,2025

[REICH-2024] REICH, C., DEBNATH, B., PATEL, D., & CHAKRADHAR, S. Differentiable jpeg: The devil is in the details. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2024

[TALEBI-2021] TALEBI, Hossein, KELLY, D., LUO, X., DORADE, I. G., YANG, F., MILANFAR, P., & ELAD,M. Better compression with deep pre-editing. IEEE Transactions on Image Processing, 2021

[VACCARO-2021] VACCARO,Federico, BERTINI,Marco, URICCHIO,Tiberio, DEL BIMBO,Alberto. Fast Video Visual Quality and Resolution Improvement using SR-UNet. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021

[ZHANG-2018] ZHANG, R.; ISOLA, P.; EFROS, A.A.; SHECHTMAN, E.; WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018